Chounta, I-A,, McLaren, B.M., Albacete, P., Jordan, P., & Katz, S. (2017). Modeling the zone of proximal development with a computational approach. In , X. Hu, T. Barnes, A. Hershkovitz, L. Paquette (Eds), *Proceedings of* the *10th International Conference on Educational Data Mining (EDM 2017)* (pp. 56-57).

Modeling the Zone of Proximal Development with a Computational Approach

Irene-Angelica Chounta, Bruce M. McLaren
Human-Computer Interaction Institute
Carnegie Mellon University
Pittsburgh PA, 15213, USA
{ichounta,bmclaren}@cs.cmu.edu

Patricia Albacete, Pamela Jordan, Sandra Katz
Learning Research and Development Center
University of Pittsburgh
Pittsburgh PA, 15260, USA
{palbacet, pjordan, katz}@pitt.edu

ABSTRACT

In this paper, we propose a computational approach to modeling the Zone of Proximal Development of students who learn using a natural-language tutoring system for physics. We employ a student model to predict students' performance based on their prior knowledge and activity when using a dialogue tutor to practice with conceptual, reflection questions about high-school level physics. Furthermore, we introduce the concept of the "Grey Area", the area in which the student model cannot predict with acceptable accuracy whether a student has mastered the knowledge components or skills present in a particular step.

Keywords

Natural-language tutoring systems, intelligent tutoring systems, student modeling, zone of proximal development

1. INTRODUCTION

Intelligent Tutoring Systems (ITSs) support students in grasping concepts, applying them during problem-solving activities, addressing misconceptions and in general improving students' proficiency in science, math and other areas [6]. ITS researchers have been studying the use of simulated tutorial dialogues that aim to engage students in reflective discussions about scientific concepts [4]. However, to a large extent, these systems lack the ability to gauge students' level of mastery over the curriculum that the tutoring system was designed to support. This is also challenging for human tutors, who do gauge the level of knowledge and understanding of their tutees to some degree, although they are poor at diagnosing the causes of student errors [3]. We argue that in order to provide meaningful instruction and scaffolding to students, a tutoring system should appropriately adapt the learning material with respect to both content and presentation. A way to achieve this is to dynamically assess students' knowledge state and needs. Human tutors use their assessment of student ability to adapt the level of discussion to the student's "zone of proximal development" (ZPD)—that is, "*the distance between the actual developmental level as determined by independent problem solving and the level of potential development as determined through problem solving under adult guidance or in collaboration with more capable peers*" [7]. Deriving ways to identify and formally describe the ZPD is an important step towards understanding the mechanisms that drive learning and development, gaining insights about learners' needs, and providing appropriate pedagogical interventions [2]. Following the practice of human tutors, we propose a computational approach to model the ZPD of students who carry out learning activities using a dialogue-based intelligent tutoring system. We employ a student model to assess students' changing knowledge as they engage in a dialogue with the system. Based on the model's predictions, we define the concept of the "Grey Area", a probabilistic region in which the model's predictive accuracy is low. We argue that this region can be used to indicate whether a student is in the ZPD. Our research hypothesis is that we can use the outcome of the student model (i.e., the fitted probabilities that predict students' performance) to model students' ZPD. To the best of our knowledge, this is a novel approach to modeling the ZPD. Even though we focus on dialogue-based tutoring systems, we expect that our approach can be generalized and extended to other kinds of ITSs.

2. METHODOLOGY

In this study, we used data collected during three previous studies with the Rimac system to train a student model and frame the proposed approach. Rimac is a web-based natural-language tutoring system that engages students in conceptual discussions after they solve quantitative physics problems [5]. Rimac's dialogues present a directed line of reasoning (DLR) where knowledge components (KCs) relate to tutor question/student response pairings. To model students' knowledge we used an Additive Factor Model (AFM) [1]. The model predicts the probability of a student completing a step correctly as a linear function of student parameters, knowledge components and learning parameters. AFM takes into account the frequency of prior practice and exposure to skills but not the correctness of responses. The dataset consists of training sessions of 291 students over a period of 4 years (2011-2015). Students worked on physics problems that explore motion laws and address 88 knowledge components (KCs). The dataset contains in total 15,644 student responses that were classified as correct or incorrect using the AFM student model.

Our research hypothesis is that we can use the fitted probabilities, as predicted by the student model, to model the ZPD. The core rationale is that if the student model cannot predict with high accuracy whether a student will answer a tutor's question correctly, then it might be the case that the student is in the ZPD. The student model provides predictions at the step level: each step consists of one question/answer exchange from the tutorial dialogue. A step may involve one or more KCs. The classification threshold (i.e., the cutoff determining whether a response is classified as correct or incorrect) is 0.5 and it was validated by the

ROC curve for the binary classifier. We expect that the closer the prediction is to the classification threshold, the higher the uncertainty of the model and thus, the higher the prediction error. Based on our hypothesis, this window of uncertainty can be used to approximately model the student's zone of proximal development. We refer to this window as the "Grey Area".

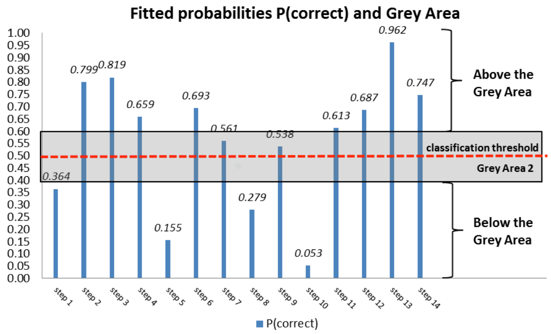

Figure 1. The Grey Area concept with respect to the fitted probabilities as predicted by the student model for a random student and for the various steps of a learning activity. Here we depict the example of a symmetrical Grey Area extending on both sides of the classification threshold.

The concept of the Grey Area is depicted in Figure 1. The space "Above the Grey Area" denotes the area where the student is predicted to answer correctly and consequently may indicate the area above the ZPD; that is, the area in which the student is able to carry out a task without any assistance. Accordingly, the space "Below the Grey Area" denotes the area where the student is predicted to answer incorrectly and consequently may indicate the area below the ZPD; that is, the area in which the student is not able to carry out the task either with or without assistance. In this paper, we model the grey area symmetrically around the classification threshold for simplicity and because the binary classifier was set to 0.5. However, the symmetry of the Grey Area is something that could change depending on the classification threshold and the learning objectives. Furthermore, we do not propose a specific size for the Grey Area. We believe that the decision about the appropriate size (or shape) of the Grey Area is not only a modeling issue but mainly a pedagogical one since it relies on the importance of the concepts taught, the teaching strategy and the learning objectives.

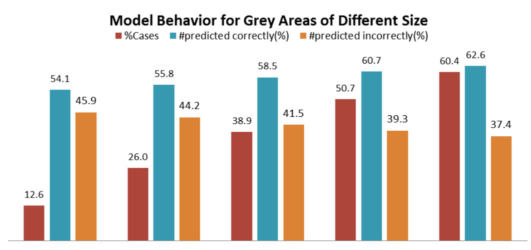

Figure 2. Model behavior (total number of predicted cases, cases predicted correctly and cases predicted incorrectly) within five grey areas of different sizes. The areas are ordered from the most narrow (Area 1) to the widest (Area 5).

Figure 2 presents an analysis of the cases that are contained in the Grey Area. In this preliminary analysis, we examined five Grey Areas of different size. On one hand, choosing a narrow grey area to model the ZPD would limit the number of cases we scaffold since fewer cases would fall within the area. On the other hand, choosing a wide grey area would affect the accuracy; that is, some cases that could be predicted correctly would be falsely labeled as "grey". However this work does not aim to define the appropriate size for the Grey Area but rather to study how the model's behavior may change for areas of different size.

3. DISCUSSION

In this paper, we present a computational approach that aims to model the Zone of Proximal Development in ITSs. To that end, we introduce the concept of the "Grey Area". Our proposal is that if the model cannot predict the state of a student's knowledge, it may be that the student is in the ZPD. We envision that the contribution of the proposed approach, besides its novelty (to the best of our knowledge there is no quantified operationalization of the ZPD) will be in defining and perhaps revising instructional methods to be implemented by ITSs. Choosing the "next step" is a prominent issue in the case of dialogue-based intelligent tutors. Not only should the task be appropriate with respect to the background knowledge of the student, but it should also be presented in an appropriate manner so that the student will not be overwhelmed and discouraged. To address this issue, we need an assessment of the knowledge state of each student and insight into the appropriate level of support the student needs to achieve the learning goals. This is described by the notion of ZPD. It is evident that if we can model the ZPD then we can adapt our instructional strategy accordingly. A limitation of our work is that we have not yet been able to conduct a rigorous evaluation of our approach; however, plans to validate our modeling methods are being developed. Our immediate plan is to carry out extensive studies to explore the proposed approach to modeling the ZPD further, as well as to better understand the strengths and limitations of using a student model to guide students through adaptive lines of reasoning.

4. REFERENCES

[1] Cen, H., Koedinger, K., and Junker, B. 2008. Comparing two IRT models for conjunctive skills. In *International Conference on Intelligent Tutoring Systems*, 796–798.

[2] Chaiklin, S. 2003. The zone of proximal development in Vygotsky's analysis of learning and instruction. *Vygotsky's educational theory in cultural context.* 1: 39–64.

[3] Chi, M.T., Siler, S.A., and Jeong, H. 2004. Can tutors monitor students' understanding accurately? *Cognition Instruct*. 22, 3: 363–387.

[4] Di Eugenio, B., Glass, M., and Trolio, M.J. 2002. The DIAG experiments: Natural language generation for intelligent tutoring systems. In *INLG02, The Third International Natural Language Generation Conference*, 120–127.

[5] Katz, S., and Albacete, P.L. 2013. A tutoring system that simulates the highly interactive nature of human tutoring. *J. Educ. Psychol.* 105, 4: 1126.

[6] VanLehn, K. 2011. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist* 46, 4: 197–221.

[7] Vygotsky, L. 1978. Interaction between learning and development. *Readings on the development of children.* 23, 3: 34–41